Networking for LHC and HEP

L. E. Price
Argonne National Laboratory

DOE/NSF Review of LHC Computing
BNL, November 15, 2000

...Thanks to much input from Harvey Newman

1

It's the Network, Stupid!

- For 20 years, high energy physicists have relied on state-of-the-art computer networking to enable ever larger international collaborations
- LHC collaborations would never have been attempted if they could not expect excellent international communications to make them possible
- The network is needed for all aspects of collaborative work
  - **Propose, design, collaborate, confer, inform**
  - **Create, move, access data**
  - **Analyze, share results, write papers**
- HEP has usually led the demand for research networks
- In special cases, we must support our own connections to high-rate locations--like CERN for LHC
  - **Because our requirements overwhelm those of other researchers**
  - **Because regional networks do not give top priority to interregional connections**

2

Networking Requirements

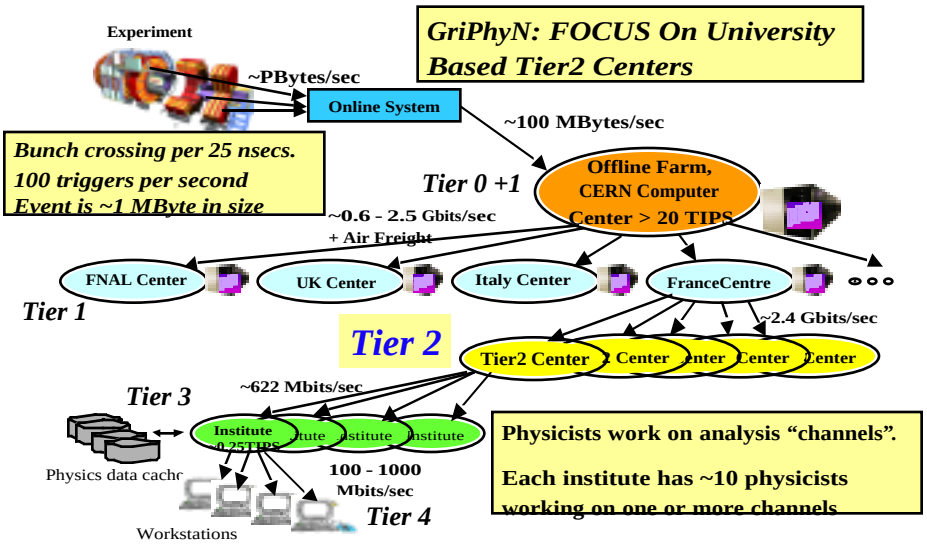
Beyond the simple requirement of adequate bandwidth, physicists in all of DoE/DHEP's (and NSF/EPP's) major programs require:

- **An integrated set of local, regional, national and international networks able to interoperate seamlessly, without bottlenecks**
- **Network and user software that will work together to provide high throughput and manage bandwidth effectively**
- **A suite of videoconference and high-level tools for remote collaboration that will make data analysis from the US (and from other remote sites) effective**

The effectiveness of U.S. participation in the LHC experimental program is particularly dependent on the speed and reliability of national and international networks

3

Networking must Support a Distributed, Hierarchical Data Access System



4

Bandwidth Requirements Projection (Mbps): ICFA-NTF

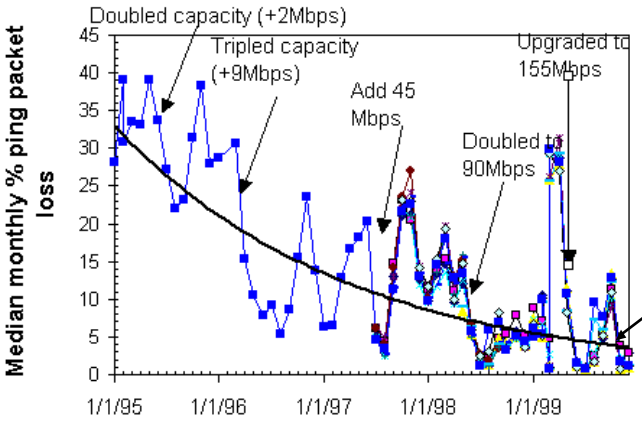
	1998	2000	>2005
BW Utilized Per Physicist (and Peak BW Used)	0.05 - 0.25 (0.5 - 2)	0.2 – 2 (2-10)	0.8 – 10 (10 – 100)
BW Utilized by a University Group	0.25 - 10	1.5 - 45	34 - 622
BW to a Home Laboratory Or Regional Center	1.5 - 45	34 - 155	622 - 5000
BW on a transoceanic Link	1.5 - 20	34 - 155	622 - 5000
BW to a Central Laboratory Housing One or More Major Experiments	34 - 155	155 - 622	2500 - 10000

10^{16} bits/π x10^{7} sec = 300 Mbs (x 8 for headroom, simulations, repeats,….)

5

Shared Internet may not be good enough!

Packet loss between ESnet & UK since 1995



Transatlantic link will probably be the thinnest connection because of cost

Sites in UK track one another, so can represent with single site
2 Beacons in UK

Indicates common source of congestion

Increased capacity by 155 times in 5 years

US-CERN Link Working Group

- DOE and NSF have requested a committee report on the need for HEP-supported transatlantic networking for LHC and…
  - **BaBar, CDf, D0, ZEUS, BTeV, etc.**
- Co-chairs: Harvey Newman (CMS), Larry Price (ATLAS)
- Other experiments are providing names of members for committee
- Hope to coordinate meeting with ICFA-SCIC (Standing Committee on Interregional networking--see below.)
- Report early in 2001.

7

Committee history: ICFA NTF

Recommendations concerning Inter-continental links:

- ICFA should encourage the provision of some considerable extra bandwidth, especially across the Atlantic
- ICFA participants should make concrete proposals, (such as recommendation to increase bandwidth across the Atlantic, approach to QoS , co-operation with other disciplines and agencies, etc.)
- The bandwidth to Japan needs to be upgraded
- Integrated end-to-end connectivity is primary requirement, to be emphasized to continental ISPs, and academic and research networks

8

ICFA Standing Committee on Interregional Connectivity (SCIC)

ICFA Commissioned the SCIC in Summer 1998 as a standing committee to deal with the issues and problems of wide area networking for the ICFA community

CHARGE

- Make recommendations to ICFA concerning the connectivity between American Asia and Europe.
- Create subcommittees when necessary to meet the charge *(Monitoring, Requirements, Technology Tracking, Remote Regions).*
- Chair of the committee should report to ICFA once per year, at its joint meeting with laboratory directors.

MEMBERSHIP

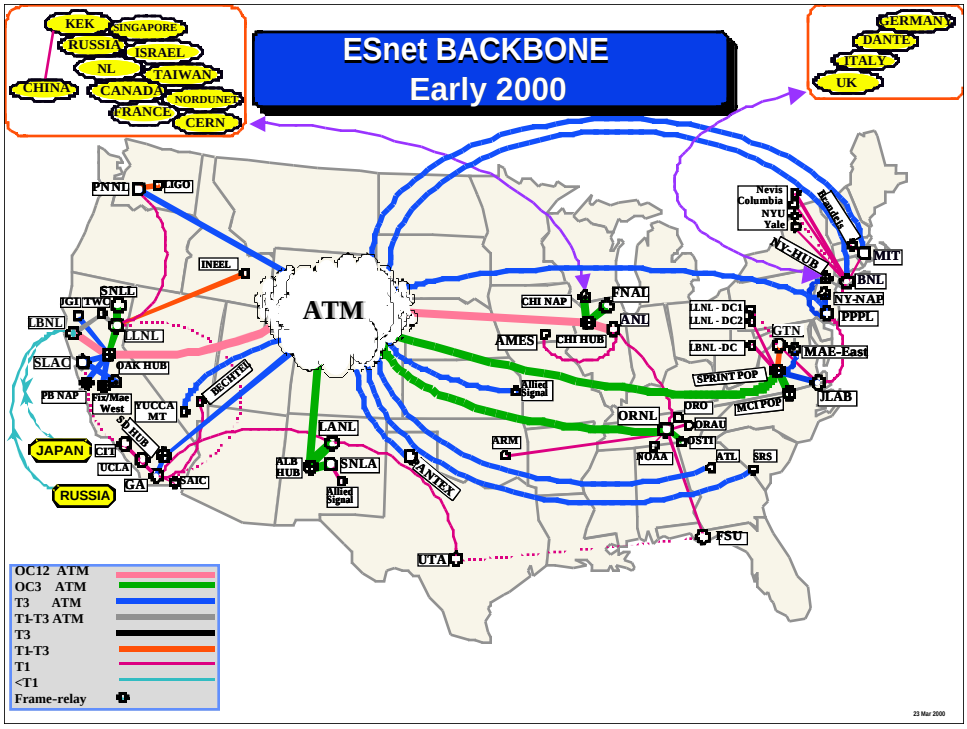
- M. Kasemann (FNAL), Chair
- H. Newman (CIT) for US Universities and APS/DPF
- Representatives of HEP Labs: SLAC, CERN, DESY, KEK
- Regional Representatives: from ECFA, ACFA, Canada, the Russian Federation, and South America

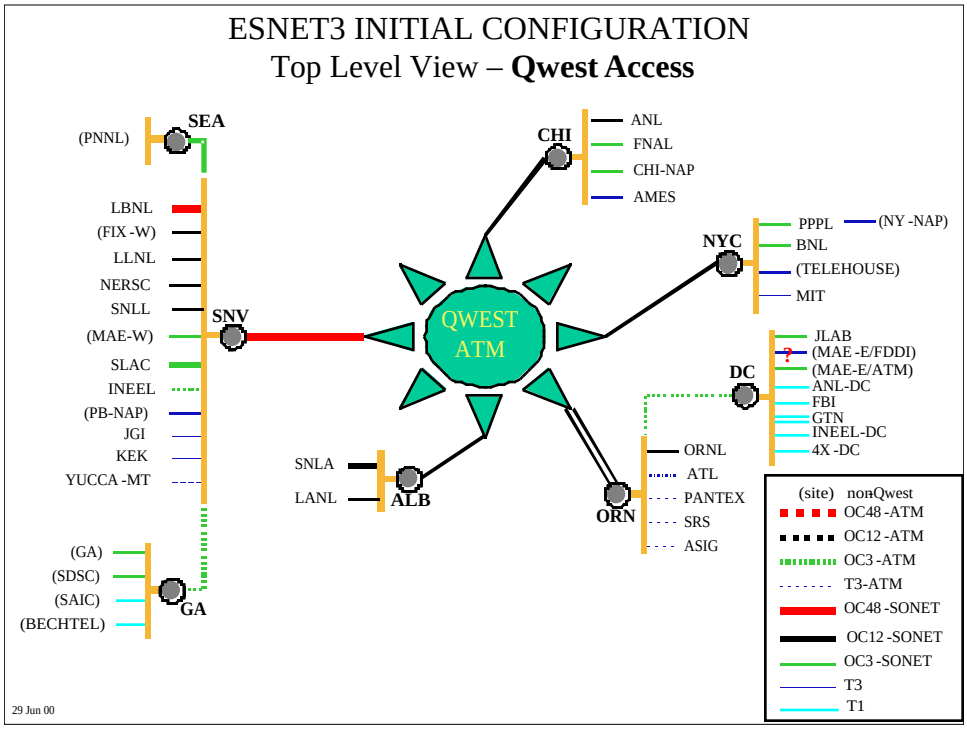
9

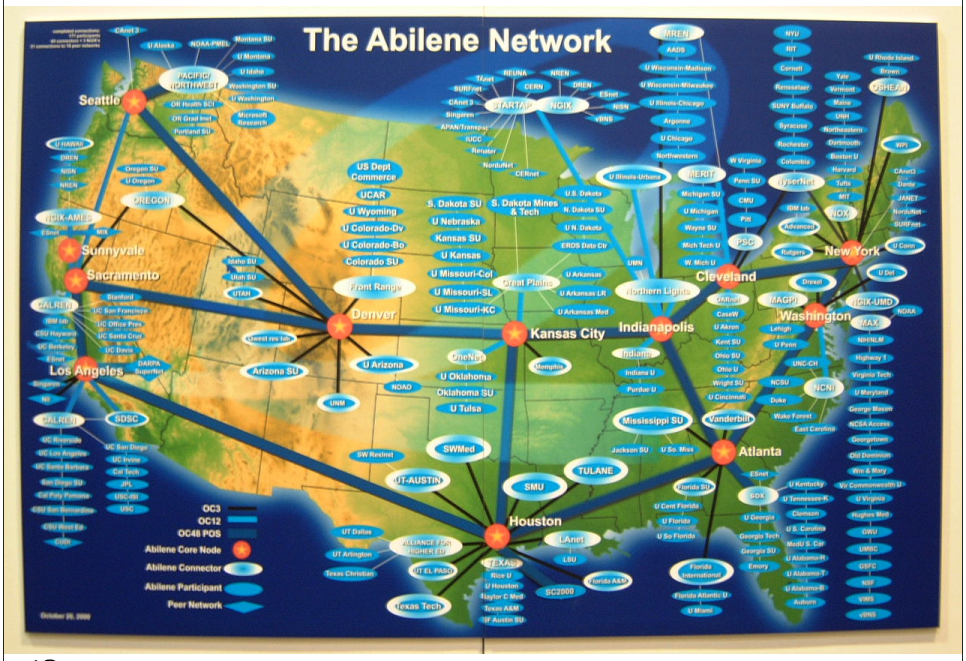
Academic & Research Networking in the US

- **Focus on research & advanced applications**
  - **hence, separate connections to commodity Internet and research backbone (GigaPoP)**
  - **lot of resistance to connect K-12 schools**
  - **Internet2 infrastructure:**
    - **vBNS**
    - **Abilene**
    - **STAR TAP**
  - **Internet2 projects:**
    - **Digital Video Initiative (DVI),**
    - **Digital Storage Infrastructure (DSI),**
    - **Qbone,**
    - **Surveyor**
- **Mission-oriented networks**
  - ***Esnet: support of Office of Science, especially Laboratories***
  - ***NASA Science Internet***

10


ESnet BACKBONE
Early 2000
KEK
SINGAPORE
RUSSIA
ISRAEL
NL
TAIWAN
CHINA
CANADA
NORDUNET
FRANCE
CERN
GERMANY
DANTE
ITALY
UK
ATM
PNNL
JGO
INEEL
SNLL
JGI
TWC
LBNL
LLNL
SLAC
OAK HUB
PB NAP
Fix/Mae West
YUCCA MT
SUNN HUB
JAPAN
CIT
UCLA
GA
SAIC
BECHTEL
RUSSIA
LANL
ALB HUB
SNLA
Allied Signal
ANTEX
UTA
CHI NAP
FNAL
ANL
AMES
CHI HUB
LLNL - DC1
LLNL - DC2
LBNL -DC
GTN
MAE-East
SPRINT POP
MCI POP
JLAB
ORNL
ORO
ORAU
OSTI
NOAA
ATL
SRS
FSU
ARM
Nevis
Columbia
NYU
Yale
Starlight
NY-HUB
MIT
BNL
NY-NAP
PPPL
OC12 ATM
OC3 ATM
T3 ATM
T1-T3 ATM
T3
T1-T3
T1
<T1
Frame-relay
23 Mar 2000


ESNET3 INITIAL CONFIGURATION
Top Level View – Qwest Access
SEA
(PNNL)
LBNL
(FIX -W)
LLNL
NERSC
SNLL
(MAE-W)
SLAC
INEEL
(PB-NAP)
JGI
KEK
YUCCA -MT
SNV
(GA)
(SDSC)
(SAIC)
(BECHTEL)
GA
CHI
ANL
FNAL
CHI-NAP
AMES
NYC
PPPL
(NY -NAP)
BNL
(TELEHOUSE)
MIT
DC
JLAB
(MAE -E/FDDI)
(MAE-E/ATM)
ANL-DC
FBI
GTN
INEEL-DC
4X -DC
QWEST
ATM
SNLA
LANL
ALB
ORN
ORNL
ATL
PANTEX
SRS
ASIG
(site) nonQwest
OC48 -ATM
OC12 -ATM
OC3 -ATM
T3-ATM
OC48 -SONET
OC12 -SONET
OC3 -SONET
T3
T1
29 Jun 00



13

Internet 2

Abilene int'l peering

SEATTLE
CAnet, (AARnet)

STAR TAP
APAN/TransPAC, Canet, IUCC, NORDUnet, RENATER, REUNA, SURFnet, SingAREN, SINET, TAnet2 (CERnet, HARnet)

OC12

NYC
DANTE*, JANET, NORDUnet, SURFnet (CAnet)

SUNNYVALE
(SINET?)

L.A.
SingAREN, (SINET?)

San Diego
(CUDI?)

El Paso, TX
(CUDI?)

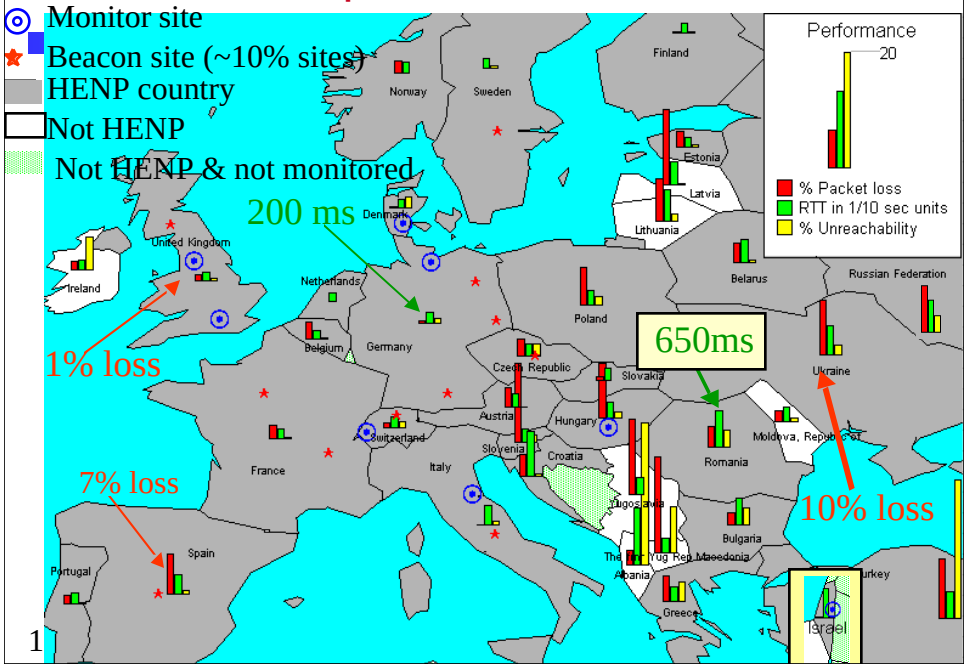
OC3-12

MIAMI
(CUDI?, REUNA, RNP2, RETINA)

14

Performance

Europe seen from U.S.



1

History of the "LEP3NET" Network (1)

Since the early days of LEP, DOE has supported a dedicated network connection to CERN, managed by Caltech

Initially dedicated to L3 experiment, more recently the line has supported US involvement in LEP and LHC

- **1982 - 1986: Use of Int'l public X.25 networks (2.4 - 9.6 kbps) to support U.S. participation in DESY and CERN programs**
- **1986 -1989: Leased analog (16.8 kbits/s) CERN-MIT X.25 switched line, with onward connections to Caltech, Michigan, Princeton, Harvard, Northeastern, ...**
- **1989 - 1991: Leased digital (64 kbits/s) CERN-MIT switched supporting L3 and also providing the US-Europe DECNET service.**
- **1991 - 1995: Leased digital (256-512 kbits/s) CERN-MIT line split to provide IP (for L3) and DECNET (for general purpose Europe-US HEP traffic)**
- **12/95 - 9/96: Major partner in leased digital (1.544Mbits/s) CERN-US line for all CERN-US HEP traffic.**
  Development of CERN-US packet videoconferencing and packet/Codec hybrid systems.

16

History of the "LEP3NET" Network (2)

October 1996 - August 1997

- Upgraded leased digital CERN-US line: **2.048 Mbps**
- Set-up of monitoring tools and traffic control
- Start Deployment of VRVS a Web-based videoconferencing system

September 1997 - April 1999

- Upgraded leased CERN-US line to **2 X 2.048 Mbps;** Addition of a backup and "overflow" leased line at **2.048 Mbps (total 6 Mbps)** to avoid saturation in Fall 1998
- Production deployment of VRVS software in the US and Europe (to 1000 hosts by 4/99; Now 2800).
- Set-up of CERN-US consortium rack at Perryman to peer with ESnet and other international nets
- Test of QoS features using new Cisco software and hardware

17

History of the "LEP3NET" Network (3)

October 1998 - September 1999

- Market survey and selection of Cable&Wireless as ISP.
- Began Collaboration in Internet2 applications and network developments.
- Move to C&W Chicago PoP, to connect to **STARTAP.**
- From April 1999, set-up of a **12 Mbps** ATM VP/VBRnrt circuit between CERN and C&W PoP
- 9/99: Transatlantic upgrade to **20 Mbps** September 1st, coincident with CERN/IN2P3 link upgrade
- 7/99: Begin organized file transfer service to "mirror" Babar DST data from SLAC to CCIN2P3/Lyon

With the close of LEP and the rise of the more demanding LHC and other programs, we are renaming the network "LHCNET"

18

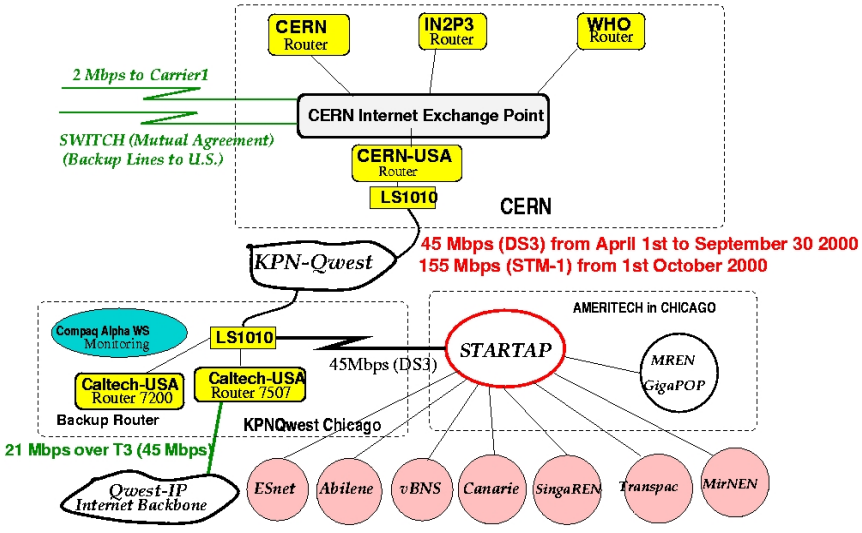
History of the “LEP3NET” Network (4)

October 1999 - September 2000

- CERN (represented by our consortium) became a member of UCAID (Internet2)
- Market survey and selection of KPN/Qwest as ISP.
- Move from C&W Chicago PoP to KPN/Qwest Chicago PoP and connection to STARTAP end of March.
- From April 2000, set-up of a 45 Mbps (DS3 SDH) circuit between CERN and KPN/Qwest PoP and 21 Mbps for general purpose Internet via QwestIP.
- October 2000: Transatlantic upgrade to 155 Mbps (STM-1) with move to the KPN/Qwest PoP in New-York with direct peering with Esnet, Abilene and Canarie (Canada).
- Possibility to have a second STM-1 (two unprotected circuits) in 2001; second one for R&D.

19

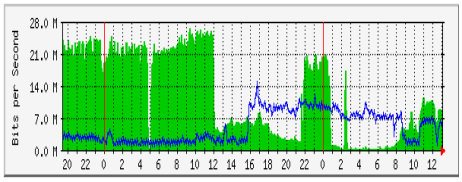
Configuration at Chicago with KPN/Qwest



20

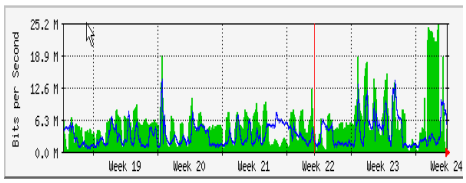
Daily, Weekly, Monthly and Yearly Statistics on the 45 Mbps line

'Daily' Graph (5 Minute Average)



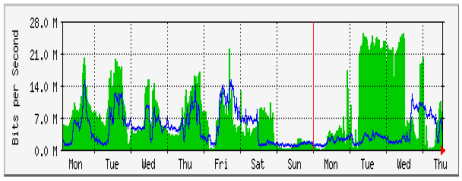
Max In: 26.8 Mb/s (59.6%) Average In: 12.3 Mb/s (27.4%) Current In: 8677.1 kb/s (19.3%)
Max Out: 15.1 Mb/s (33.5%) Average Out: 5150.2 kb/s (11.4%) Current Out: 6704.4 kb/s (14.9%)

'Monthly' Graph (2 Hour Average)



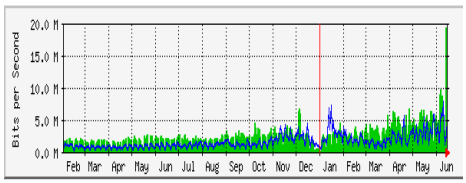
Max In: 25.1 Mb/s (55.9%) Average In: 5612.4 kb/s (12.5%) Current In: 4584.4 kb/s (10.2%)
Max Out: 14.2 Mb/s (31.5%) Average Out: 3531.2 kb/s (7.8%) Current Out: 3889.4 kb/s (8.6%)

'Weekly' Graph (30 Minute Average)



Max In: 25.6 Mb/s (57.0%) Average In: 8255.1 kb/s (18.3%) Current In: 8500.8 kb/s (18.9%)
Max Out: 15.2 Mb/s (33.8%) Average Out: 4594.6 kb/s (10.2%) Current Out: 5346.9 kb/s (11.9%)

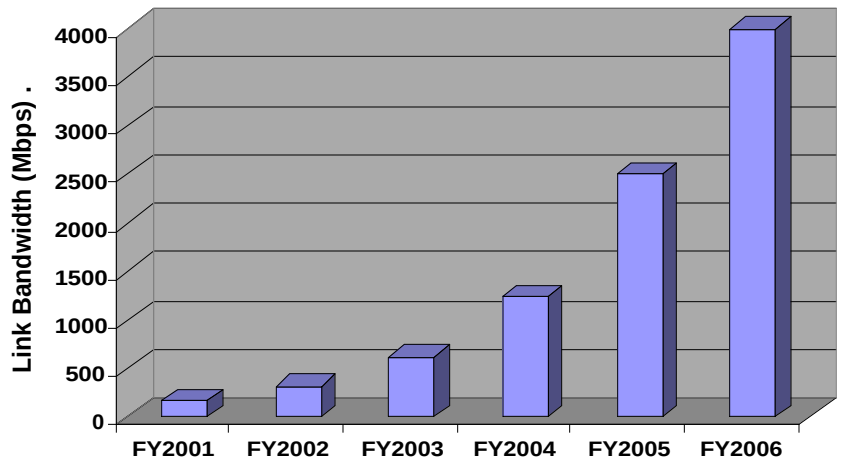
'Yearly' Graph (1 Day Average)



Max In: 19.5 Mb/s (43.4%) Average In: 2582.0 kb/s (5.7%) Current In: 13.5 Mb/s (30.1%)
Max Out: 7997.9 kb/s (17.8%) Average Out: 1726.7 kb/s (3.8%) Current Out: 5333.6 kb/s (11.9%)

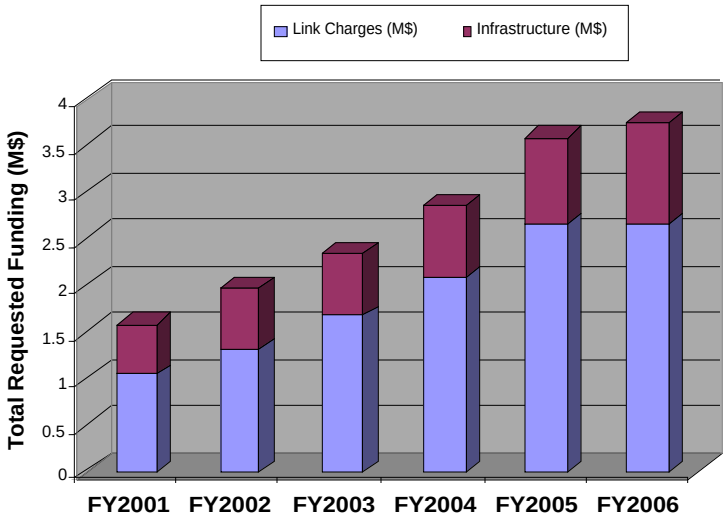
21

Bandwidth Requirements for the Transatlantic Link



22

Estimated Funding for Transatlantic Link



23

CERN

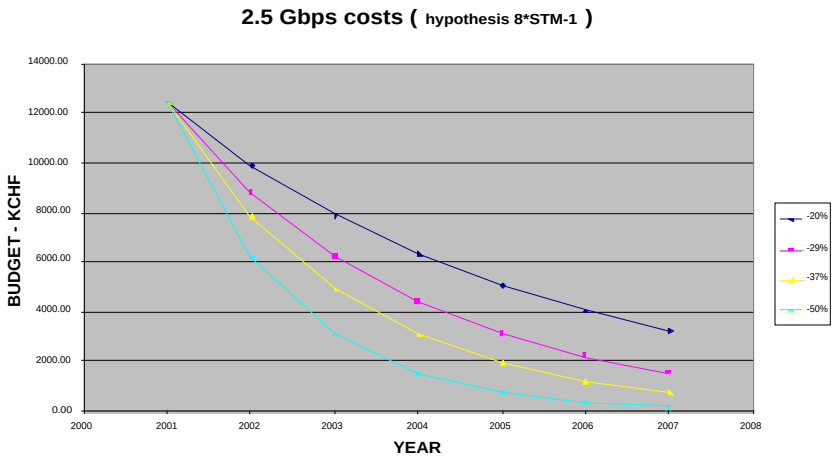
Unit Costs are Going Down

Recent price history on CERN-US link:

- **still paying 400KCHF/Mbps/year 16 months ago (Swisscom/MCI),**
- **then 88KCHF/Mbps/year (C&W)**
- **now 36KCHF/Mbps/year (KPN-Qwest)**
- **expect to pay 8KCHF/Mbps/year, if the dual unprotected STM-1 solution is selected.**

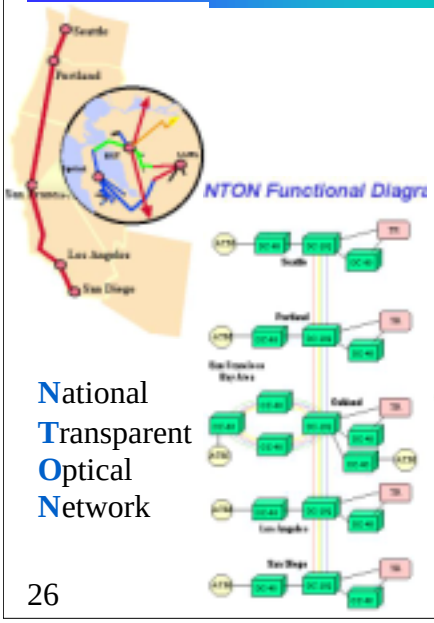
24

2.5 Gbps scenarios



25

We are Preparing Now to Use this Large Bandwidth When it Arrives



National
Transparent
Optical
Network

Nov. 9, 2000 at SC2000:

- **Peak transfer rate of 990 Mbs measured in test from dallas to SLAC via NTON**
- **Best results achieved with 128KB window size and 25 parallel streams**
- **Demonstration by SLAC and FNAL of work for PPDG**

Caltech and SLAC working toward 2Gbs transfer rate over NTON in 2001

Need for differentiated services (QoS)

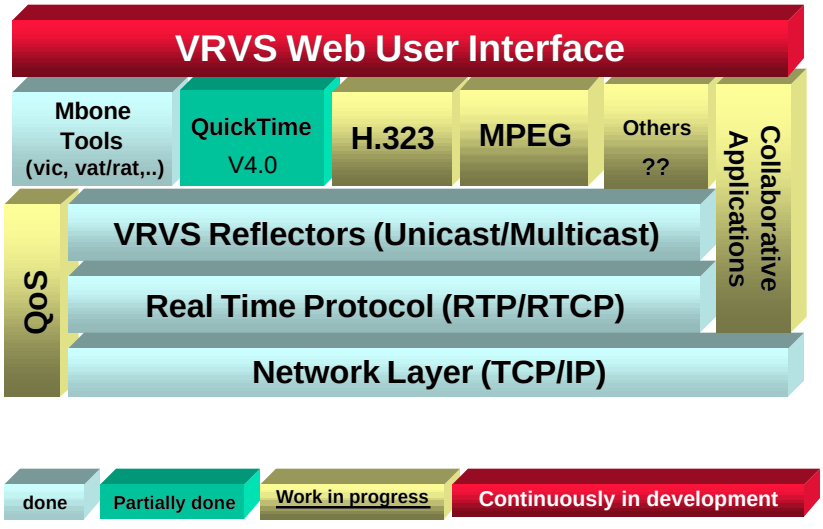
26

Network must also support advanced conferencing services: *e.g.*, VRVS

Example: ***9 Participants, CERN(2), Caltech, FNAL(2), Bologna (IT), Roma (IT), Milan (IT), Rutherford(UK)***

27

Continued Development of VRVS



28

Adequate networking for LHC turnon is only the start!

A Short List of Coming Revolutions

Network Technologies

- ***Wireless Broadband (from ca. 2003)***
- ***10 Gigabit Ethernet (from 2002: See www.10gea.org)***
  10GbE/DWDM-Wavelength (OC-192) integration: OXC

Internet Information Software Technologies

- **Global Information "Broadcast" Architecture**
  - ***E.g the Multipoint Information Distribution Protocol (MIDP; Tie.Liao@inria.fr)***
- **Programmable Coordinated Agent Archtectures**
  - ***E.g. Mobile Agent Reactive Spaces (MARS) by Cabri et al., Univ. of Modena***

The "Data Grid" - Human Interface

- **Interactive monitoring and control of Grid resources**
  - ***By authorized groups and individuals***
  - ***By Autonomous Agents***

29

CERN

WAN vs LAN bandwidth

The common belief that WAN will always be well behind LANs (i.e. 1-10%) may well be plain wrong….

- ***WAN technology is well ahead of LAN technology, state of the art is 10Gbps (WAN) against 1Gbps (LAN)***
- ***Price is less of an issue as they are falling down very rapidly.***
- ***Some people are even advocating that one should now start thinking new applications as if bandwidth was free, which sounds a bit premature to me, at least, in Europe, even though there are large amounts of unused capacity floating around!***

30

Conclusions

- Seamless high-performance network will be crucial to success of LHC--and other international HEP experiments
  - **Data transfer and remote access**
  - **Rich menu of collaboration and conferencing functions**
- We are only now realizing that networking must be planned as a large-scale priority task of major collaborations--it will not automatically be there
  - **BaBar is scrambling to provide data transport to IN2P3 and INFN**
- Advance of technology means that the networking we need will not be as expensive as once feared.
  - **But *a fortiori* we should not provide less than we need**
- The US-CERN Link Working Group will have an interesting and vital task
  - **Evaluate future requirements and opportunities**
  - **Recommend optimum cost/performance tradeoff**
  - **Pave the way for effective and powerful data analysis**

31